On the number of distinct elements in a random sample drawn with replacement from a finite population.

BU-250-M D.S. Robson and B.J. Rothschild December, 1967

ABSTRACT

A sequence of independent random numbers generated by a discrete uniform probability distribution on the integers 1,2,...,R is commonly employed as a device for drawing random samples. If a sample of size $\underline{s}$ is to be drawn without replacement from a finite population of size $N \leq R$ then the probability that at least $\underline{m}$ random numbers will be required to obtain $\underline{s}$ distinct integers from the set $\{1,2,\ldots,N\}$ is

$$F_s(m) = 1 - \binom{N-1}{s-1} \sum_{\nu=0}^{s-1} \binom{s-1}{\nu} (-1)^{s-1-\nu} \left(\frac{N}{N-\nu}\right)\left(1 - \frac{N-\nu}{R}\right)^m.$$

This result has direct application to sequential mark-recapture theory when individuals are captured singly, marked, and replaced into the population.

On the number of distinct elements in a random sample drawn with replacement from a finite population.

BU-250-M D.S. Robson and B.J. Rothschild December, 1967

Introduction

The mechanics of drawing a simple random sample without replacement from a finite population of size N commonly consist first of numbering the population elements from 1 to N and then drawing a sequence of numbers from a random number table. This sequence will in general contain both usable and unusable random numbers; a number is unusable if and only if it either falls outside the range 1 to N or falls within this range but is a repeat of an earlier number in the sequence. We shall suppose that the range of possible random numbers is from 1 to R, $R \geq N$, so the numbers N + 1 to R are unusable.

The length M of the random number sequence required to obtain a sample of fixed size $\underline{s}$ drawn without replacement from the population $J_N = \{1,2,\cdots,N\}$ is then a random variable having a probability distribution $F_s(m)$ depending on N,R and s. Similarly, the number S of distinct numbers from J_N which appear in a random number sequence of fixed length $\underline{m}$ is a random variable having a probability distribution $H_m(s)$ depending on N,R and m. Since the event "at most $\underline{m}$ random numbers are required to produce exactly $\underline{s}$ distinct numbers from J_N" is equivalent to the event "at least $\underline{s}$ distinct numbers from J_N appear among the first $\underline{m}$ random numbers" then the conditional distribution functions $F_s(m)$ and $H_m(s)$ are related by

$$(1) \quad F_s(m) = P(M \leq m | s) = P(S \geq s | m) = 1 - H_m(s-1)$$

We shall determine these distributions explicitly by calculating $H_m(s)$.

The distribution of sample sizes generated by random number sequences of fixed length.

Let S_m denote the number of distinct elements of J_N appearing among the first m elements of a random number sequence, then S_m is stochastically related to S_{m-1} by

$$(2) \qquad S_m = \begin{cases} S_{m-1} \text{ with probability } 1 - \dfrac{N-S_{m-1}}{R} \\ S_{m-1} + 1 \text{ with probability } \dfrac{N-S_{m-1}}{R} \end{cases}$$

It follows that the factorial moments of S_m may be determined from

$$E\left[S_m^{(k)} \right] = E\left[S_{m-1}^{(k)} \left(1 - \frac{N-S_{m-1}}{R} \right) + \left(S_{m-1} + 1 \right)^{(k)} \left(\frac{N-S_{m-1}}{R} \right) \right]$$

$$= \left(1 - \frac{k}{R}\right) E\left(S_{m-1}^{(k)}\right) + \frac{k(N-k+1)}{R} E\left(S_{m-1}^{(k-1)}\right)$$

by solving recursively to obtain

$$E\left[S_m (S_m-1) \cdots (S_m-k+1) \right] = N^{(k)} \sum_{\nu=0}^{k} \binom{k}{\nu} (-1)^{\nu} \left(1 - \frac{\nu}{R}\right)^m$$

where $x^{(k)} = x\,(x-1) \cdots (x-k+1)$.

Similarly, letting

$$h_m(s) = P(S = s|m)$$

we have from (2)

$$h_m(s) = \frac{N - s + 1}{R} h_{m-1}(s-1) + (1 - \frac{N - s}{R}) h_{m-1}(s)$$

giving

$$h_m(s) = \binom{N}{s} \sum_{\nu=0}^{s} \binom{s}{\nu} (-1)^{s-\nu} \left(1 - \frac{N-\nu}{R}\right)^m$$

and

$$(3) \qquad H_m(s) = \sum_{i=0}^{s} h_m(i) = \binom{N-1}{s} \sum_{\nu=0}^{s} \binom{s}{\nu} (-1)^{s-\nu} \left(\frac{N}{N-\nu}\right) \left(1 - \frac{N-\nu}{R}\right)^m$$

<u>The distribution of the length of random number sequences producing samples of fixed size</u>.

Having calculated $H_m(s)$ we may apply (1) to obtain $F_s(m)$, and letting

$$f_s(m) = P(M = m|s)$$

we have

$$f_s(m) = H_{m-1}(s-1) - H_m(s-1)$$

$$= \frac{N - s + 1}{R} h_{m-1}(s-1)$$

Notice that the latter result may be obtained directly from the argument that the m'th random number yields the s'th distinct number from J_N if and only if there were s-1 distinct numbers from J_N among the first m-1 random numbers <u>and</u> the m'th random number is one of the N - (s-1) members of J_N which had not previously appeared.

The probability generating function is then

$$E\left[\, t^m | s \,\right] = st \binom{N}{s} \sum_{\nu=0}^{s-1} \binom{s-1}{\nu} (-1)^{s-1-\nu} \left[\frac{1}{R(1-t) + (N-\nu)} \right]$$

and the factorial moments of M are

$$E\left[M^{(k)} | s \right] = k! \, \frac{s}{R} \binom{N}{s} \sum_{\nu=0}^{s-1} \binom{s-1}{\nu} (-1)^{s-1-\nu} \left(\frac{R}{N-\nu} \right)^{k+1} \left(1 - \frac{N-\nu}{R} \right)^{k-1} .$$

In particular, the mean value of M may be expressed as

$$E\left[M | s \right] = \frac{R}{N} \sum_{\nu=1}^{s} \frac{\binom{s}{\nu}}{\binom{N-1}{\nu-1}}$$

An application to the mark-recapture problem.

Let R = N denote the unknown number of animals in a population and suppose that animals are captured singly (at random), marked, and replaced into the population. The number S of different animals captured in a fixed number m of trials is then a sufficient statistic with respect to the unknown population size N, and the probability distribution of S is, from (3),

$$P(S \le s | m) = \binom{N-1}{s} \sum_{\nu=0}^{s} \binom{s}{\nu} (-1)^{s-\nu} \left(\frac{N}{N-\nu} \right) \left(\frac{\nu}{N} \right)^m .$$